Data Scientist Coding Challenge

Thank you for your interest in the Data Scientist role at Plaid! As a next step, we would like you to complete a short coding challenge to get a sense of how you approach data science related problems. Please complete this coding challenge in the next few days.

Challenges

This challenge comes with two csv files, ​code\_challenge\_train.csv ​and ​code\_challenge\_test.csv​. Each row in these two csv files contains a pair of financial securities, which are represented by an id and a description. Two securities are considered to be identical if they have the same id, even if the descriptions don't exactly match.

Your challenge is to predict which rows in the test dataset represent identical securities, using only the security descriptions as features. The difficulty of predicting each row will vary significantly, so please do not aim for 100% accuracy. There are several good ways to approach this, and we have no preference between them.

Additional Instructions

Please keep all code and other project files in a directory labeled with your full name. Submit this directory as a zip (or rar, 7z, tar, etc.) file.

Lastly, we’re not looking for perfection. We want you to have fun coding and show us your skills. Feel free to reach out if you have any questions!

Best,

Plaid Team